Is één klas voldoende? De betrouwbaarheid van leerlingenobservaties opnieuw bekeken met oog op het risico van “high-stake” besluiten

van der Lans, Rikkert; Helms-Lorenz, Michelle; Maulana, Ridwan

Is één klas voldoende? De betrouwbaarheid van leerlingenobservaties opnieuw bekeken met oog op het risico van "high-stake" besluiten

Deze studie focust op de betrouwbaarheid van leerlingenobservaties voor docenten in het voortgezet onderwijs (VO). Uit eerder onderzoek naar de betrouwbaarheid van studentenobservaties in hoger onderwijs (HO) blijkt dat studenten hun docenten met hoge betrouwbaarheid kunnen evalueren, maar dit resultaat is nog niet gerepliceerd met leerlingen in het VO. Ook is er kritiek op de onderzoeksopzet voor de bepaling van de betrouwbaarheid. Deze studie verkent daarom de betrouwbaarheid van leerlingenobservaties van docenten VO nagegaan mbv twee verschillende onderzoeksopzetten: een geneste en een gekruiste. De steekproef bestond uit 407 leerlingen uit 17 klassen. De resultaten bevestigen de betrouwbaarheid van leerlingenobservaties, maar suggereren tegelijk dat reguliere schattingen van de betrouwbaarheid optimistischer zijn dan schattingen die gebaseerd zijn op een complexere gekruiste onderzoeksopzet.

Introductie

Deze studie is gemotiveerd uit de beleidsagenda om de kwaliteit van lesgeven meer transparant te maken en individuele docenten in grotere mate aansprakelijk te stellen voor geleverde kwaliteit (NCTQ, 2013; Nusche, et al. 2014). Dit beleid zorgt ervoor dat de aard, de frequentie en het belang van onderwijsevaluatie aan het veranderen is (Marzano & Toth, 2013). In Nederland wordt er bijvoorbeeld naar gestreefd dat in 2020 100% van de docenten primair onderwijs (PO) en voortgezet onderwijs (VO) jaarlijks functioneringsgesprekken krijgt, iets waar Nederland historisch geen traditie in heeft (Nusche, et al. 2014).

In relatief korte tijd zijn ook de adviezen voor implementatie en gebruik van instrumenten steeds verder gewijzigd. Oorspronkelijk was het idee dat evaluatie voor "high-stake" besluiten zou plaatsnemen op basis van gestandaardiseerde leerlingenprestatietoetsen en evaluatie voor "low-stake" feedback op basis van lesobservaties en leerlingenobservaties (zie bijvoorbeeld Kane, et al. 2012). Maar recent worden lesobservaties en leerlingenobservaties over het pedagogisch-didactisch handelen steeds vaker geopperd in de context van high-stake besluiten (Darling-Hammond, 2013; Marzano & Toth, 2013; NCTQ, 2013). Eén van de hoofdredenen hiervoor is de (te) lage betrouwbaarheid van gestandaardiseerde leerlingenprestatietoetsen.

Bovengenoemde ontwikkelingen vergroten de kans dat leerlingenobservaties (in de toekomst) worden meegenomen in "high-stakes" besluiten over aanstellingen en/of salarisverhoging. Echter, in tegenstelling tot lesobservaties, is over leerlingenobservaties geringe kennis over de betrouwbaarheid voor high-stake gebruik. Eerdere studies naar de betrouwbaarheid van leerlingenobservaties komen voornamelijk uit het hoger onderwijs (HO) (e.g. Marsh, 2007) en kennen meestal eenzelfde onderzoeksopzet (Morley, 2012). Morley argumenteert dat deze veel toegepaste onderzoeksopzet varianties verstrengelt, waardoor de betrouwbaarheid overschat wordt.

De studie stelt zich daarom twee doelen. De eerste is om de eerdere bevindingen uit het HO te repliceren in de context van het VO; de tweede is om te verkennen of er aanleiding is te suggereren dat de betrouwbaarheid van leerlingenobservaties wordt overschat.

Onderzoeksvragen:

1. Is de betrouwbaarheid van leerlingenobservaties in het VO vergelijkbaar met de betrouwbaarheid van studentenobservaties in het HO, zoals gerapporteerd in Marsh (2007)?
2. Zijn er aanwijzingen dat eerdere studies de betrouwbaarheid van leerlingenobservaties hebben overschat?

Methode

Steekproef

De steekproef voor dit onderzoek bestond uit 409 leerlingen uit 17 klassen uit de onderbouw VO (leeftijd 12 tot 15 jaar). De data komt van acht scholen. Iedere klas observeerde de kwaliteit van het pedagogisch-didactisch handelen van 3 of 4 docenten. In totaal deden 63 docenten mee aan het onderzoek.

Instrument

De gebruikte Mijn Leraar vragenlijst telt 40 items die tezamen het pedagogisch-diactisch handelen van docenten evalueren (Maulana, et al. 2015).

Analyse

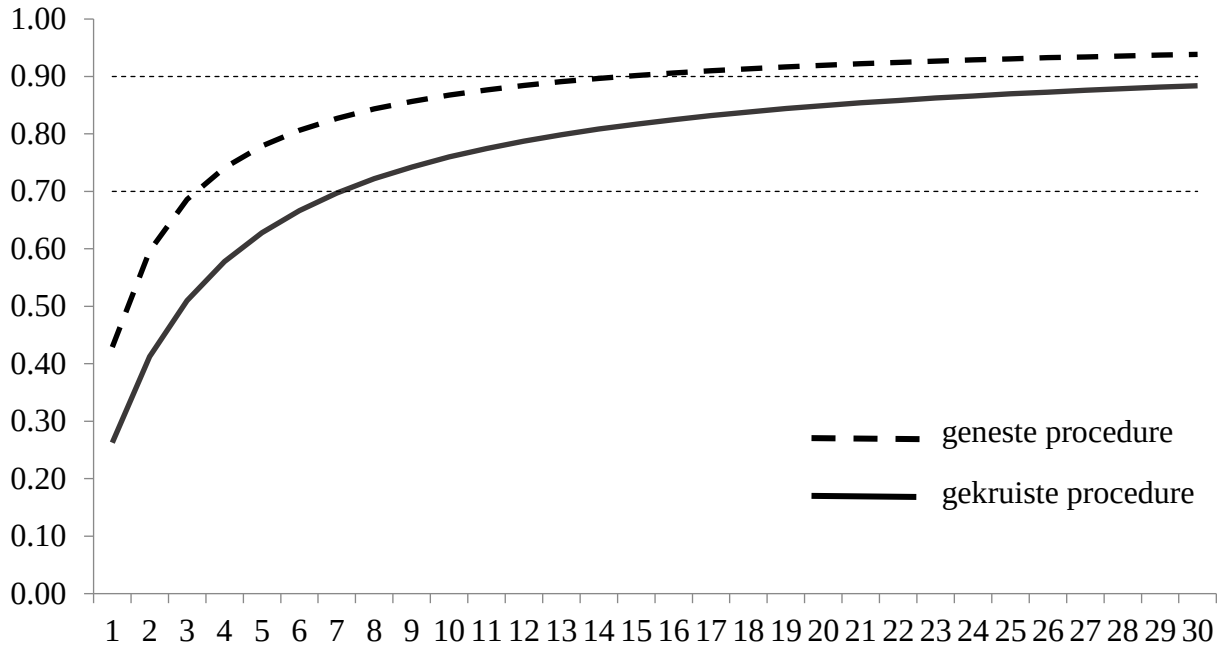
De analyse maakt gebruik van het Geraliseerbaarheid in Item Response Theorie (GIRT) raamwerk (Choi, 2013). Om de analyses genoemd door Marsh (2007) te repliceren in het VO is van iedere leerling willekeurig één van de drie of vier beschikbare leerlingenobservaties geselecteerd voor analyse. Dit resulteerde in een geneste steekproef van 409 leerlingen waarbij iedere leerling één docent evalueerde.

Morely (2012) argumenteert dat een gekruiste onderzoeksopzet, zoals één waarin leerlingen meerdere docenten beoordelen, tot meer accurate en waarschijnlijk ook lagere schatting zou komen in vergelijking met de geneste procedure. Om deze claim te onderzoeken is een analyse gedaan met een gekruiste opzet waarin van iedere leerling observaties van meerdere docenten werden meegewogen.

Resultaten en Conclusies

De Figuur 1 geeft de betrouwbaarheid aan van leerlingenobservaties. De figuur laat zien dat de betrouwbaarheid toeneemt wanneer observaties van een groter aantal leerlingen wordt samengenomen. Vergeleken met studies uit het HO, neemt de betrouwbaarheid van observaties van leerlingen VO sneller toe. De indicatie is dat bij ongeveer 15 leerlingen (in plaats van 24 leerlingen voor HO (Marsh, 2007)) betrouwbaarheid hoger is dan .90.

De analyses suggereren tegelijk dat Morely (2012) terechte kritiek heeft geuit. De schatting van de betrouwbaarheid is inderdaad lager wanneer gebruik wordt gemaakt van een gekruiste procedure. Op basis van een gekruiste onderzoeksopzet is de inschatting dat meer 37 leerlingen nodig zijn om tot het criterium van .90 te komen.



Figuur 1. *Betrouwbaarheid van leerlingenobservaties bij een toenemend aantal leerlingen.*

Implicaties

De resultaten suggereren dat leerlingenobservaties uit één klas te weinig betrouwbare informatie bieden voor high-stake besluiten. Op basis van deze data is de inschatting dat 37 leerlingen (twee klassen) voldoende betrouwbare scores leveren voor high-stake beslissingen. Een kanttekening hierbij is dat ook in de hier toegepaste gekruiste onderzoeksopzet niet alle verstrengelingen van varianties ontrafelt, waardoor ook in deze opzet de betrouwbaarheid mogelijk nog wordt overschat. Het voorlopig advies is om leerlingenobservaties te combineren met meerdere lesobservaties van verschillende personen bij high-stake besluiten.

Referenties

Choi, J. (2013). *Advances in combining generalizability theory and item response theory.* Doctoral dissertation, University of California, Berkeley.

Darling-Hammond, L. (2013). *Getting teacher evaluation right. What really matters for effectiveness and improvement.* New York, USA: Teachers College Press

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference. A new model for teacher growth and student achievement.* Alexandria, Virginia: ASCD

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*(2), 169-194.

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation*, *38*(1), 15-20.

NCTQ (2013). *Connect the dots: Using evaluations of teaching effectiveness to inform policy and practice.* Washington, DC: National Council on Teacher Quality.

Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014.* OECD Reviews of Evaluation and Assessment in Education, OECD Publishing. http://dx.doi.org/10.1787/9789264211940-en